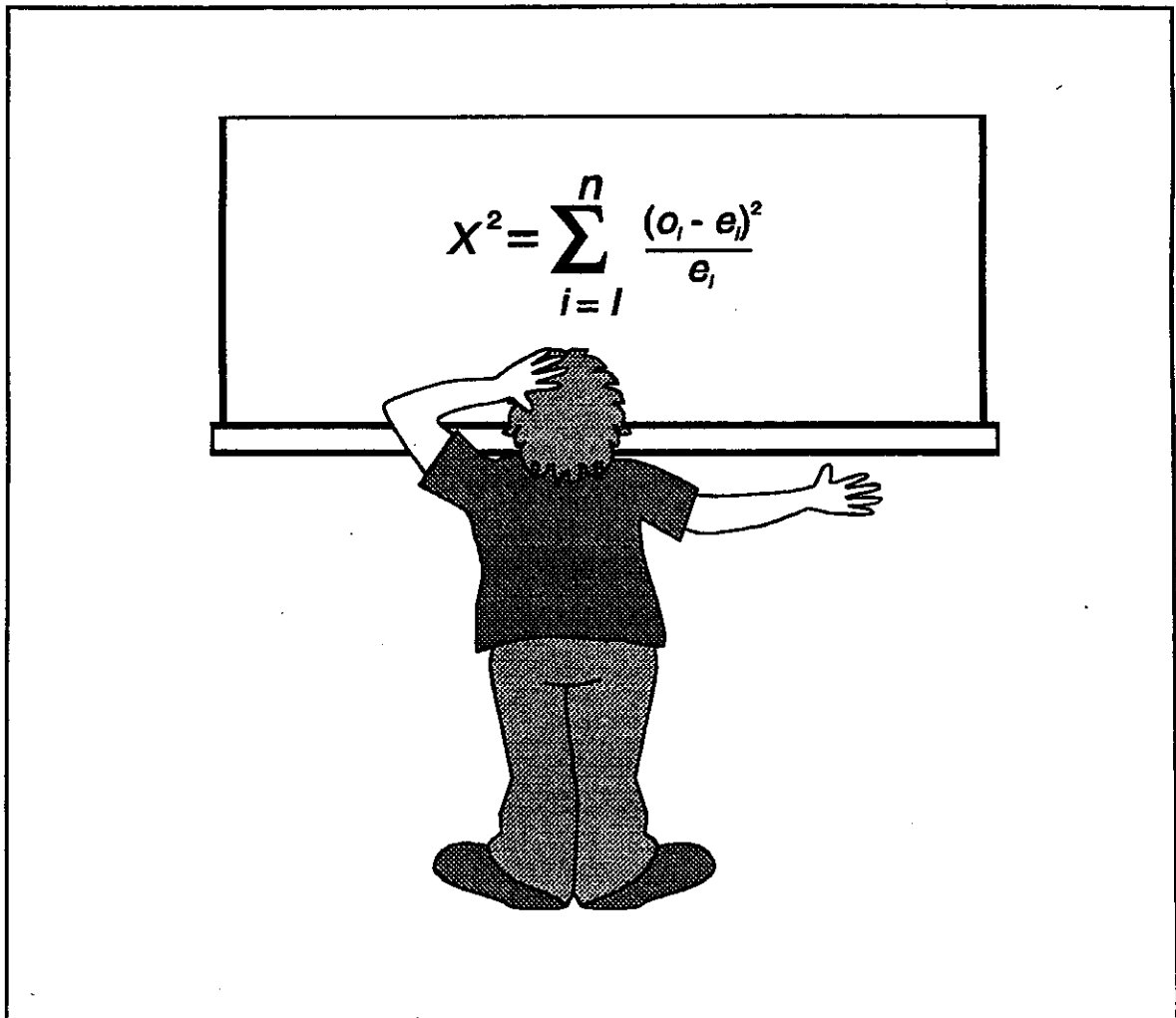


Lesson 13

The Chi-Square Distribution



Questions To Consider

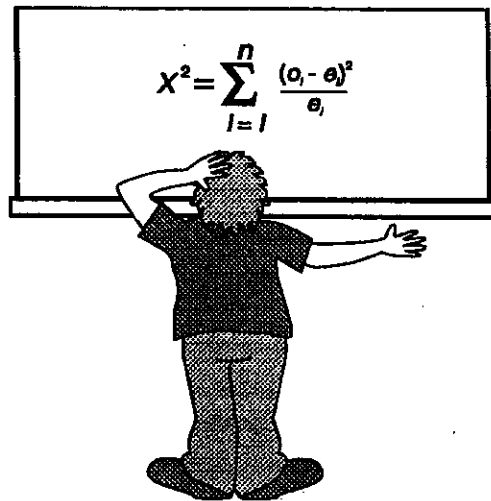
- What is a chi-square test and when is it used?
- How do you test for compatibility between observed and expected frequencies?
- How do you define a critical region for a chi-square test?
- What are degrees of freedom?
- What is a contingency table?

Key Terms

- Chi-square (X^2)
- Contingency table
- Degrees of freedom

Chi-Square

In Chapter 11, Testing of Hypotheses, you worked with the basic assumption that you knew the way the population was distributed (e.g., that it was normally distributed). From there you went on to test some assumptions about a parameter of that distribution. Well, sometimes you either don't know how the population is distributed or know that it's not distributed in such a way that you can use the methods discussed in the preceding chapters. In such cases you must use a non-parametric test. The chi-square test is such a test.



The chi-square test is used to determine whether a sample conforms to an expected standard.

The **chi-square** (X^2) distribution is a probability distribution. Just as a normal distribution can be represented graphically (recall the bell-shaped curves), a chi-square distribution, too, can be represented graphically. You'll learn more about that later, though. First, let's look at the basic idea behind a X^2 test.

You can use the X^2 test to determine whether a sample from a population conforms to an expected standard. As with other tests, you can determine whether variations from the standard are attributable merely to chance or are attributable to something else.

If the X^2 value is outside the critical value, you can accept the hypothesis; otherwise, you can reject the hypothesis.





Basically, you calculate the X^2 value and then determine whether it exceeds a predetermined ("critical") value. In other words, you determine whether or not it falls within a critical region (Recall those from Chapter 11?). If the value falls outside the critical region, you can assume that any variation is attributable to chance, and you can accept the hypothesis you are testing. If it falls within the critical region, you can reject the hypothesis.

Let's look at how this works. Suppose you perform a sampling experiment in which there are a finite number of possible outcomes. For example, suppose you have a regular deck of 52 cards. The experiment you are performing is to draw a card from the deck and record whether it is a heart, diamond, spade, or club. You repeat this 32 times, each time replacing the card just drawn. There are, of course, four possible outcomes (each of the four suits), and each has a $13/52$, or $1/4$, probability of appearing on any one draw. You would, therefore, expect that in 32 trials, you would get 8 hearts, 8 diamonds, 8 spades, and 8 clubs.

If, after 32 trials, you had 7 hearts, 10 diamonds, 9 spades, and 6 clubs, would you be able to assume that it was a regular deck? (Assume that the person drawing the cards was doing so honestly.) In other words, can you accept the hypothesis that the deck was "honest"?

If you recorded your results in a table, labeling the expected outcomes (8 of each suit) with the symbol e_i and labeling the actual, or observed, outcomes with the symbol o_i , you'd get Table 13-1.

Table 13-1. Observed And Expected Frequencies For Card-Drawing Experiment

	 $i=1$	 $i=2$	 $i=3$	 $i=4$
o_i	7	10	9	6
e_i	8	8	8	8

Now, you need to test whether the observed outcomes were close enough to the expected outcomes that you can accept the hypothesis. That is, you need to test how compatible the observed and expected frequencies are. The measure of this compatibility, or closeness, is called chi-square and can be calculated using the following formula:

$$X^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

Here is the formula for calculating chi-square.

Where: o_i = the observed frequency
 e_i = the expected frequency
 n = the number of possible outcomes

If you apply this formula to the data in Table 13-1, you get:

$$\begin{aligned} X^2 &= \frac{(7-8)^2}{8} + \frac{(10-8)^2}{8} + \frac{(9-8)^2}{8} + \frac{(6-8)^2}{8} \\ &= \frac{(-1)^2}{8} + \frac{(2)^2}{8} + \frac{(1)^2}{8} + \frac{(-2)^2}{8} \\ &= \frac{1}{8} + \frac{4}{8} + \frac{1}{8} + \frac{4}{8} \\ &= \frac{10}{8} \\ &= 1.25 \end{aligned}$$

You can tell from looking at the formula that if the observed outcomes are the same as the expected outcomes, then $X^2 = 0$. By way of illustration, suppose that $e_i = 5$ for each of three possible outcomes. If $o_i = 5$ for all three, then:

$$\begin{aligned} X^2 &= \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5} \\ &= \frac{0^2}{5} + \frac{0^2}{5} + \frac{0^2}{5} \\ &= \frac{0}{5} + \frac{0}{5} + \frac{0}{5} \\ &= 0 \end{aligned}$$

Thus, you can see that a X^2 value “close” to zero indicates a high degree of compatibility between expected and observed frequencies, and that a X^2 value “far” from zero indicates a low degree of compatibility. The trick, of course, is to find the value above which X^2 is too “far” from zero for you to be able to accept the hypothesis. In other words, you need to define a critical region for the test.

Before you do that, however, you need to be familiar with the concept of **degrees of freedom**.

To find the degree of freedom, subtract the number of outcomes considered from the total possible outcomes.

If you were told that 86 out of 100 tosses of a die were 1's, 2's, 3's, 4's, and 5's, then you'd be able to tell how many 6's showed up ($100 - 86 = 14$); in this case the experimental outcome has $v = 6 - 1 = 5$ degrees of freedom. The shape of the distribution of X^2 changes with v just as the shape of the normal distribution changes with the square of the standard deviation of the population (σ^2).

Figure 13-1 shows several different X^2 distributions. Note that the shape of the curve depends on the number of degrees of freedom.

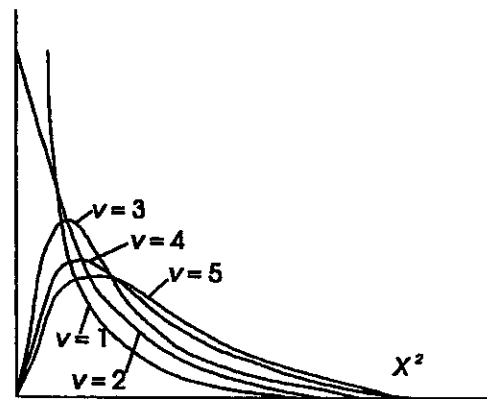
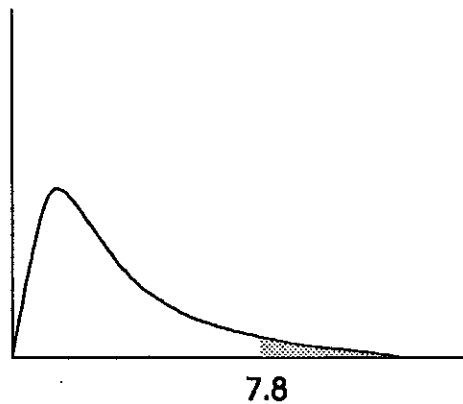


Figure 13-1. X^2 Distribution For Up To Five Degrees Of Freedom

OK, back to the problem about the cards. The X^2 distribution in Figure 13-2 is for three degrees of freedom. (Remember: you had four possible outcomes, so $v = 4 - 1 = 3$.)



The critical region is the set of values in the right tail or the left tail of a distribution.

Figure 13-2. χ^2 Distribution For Three Degrees Of Freedom

The value that marks off the 5% right tail of this distribution is 7.8. (This value was obtained from a χ^2 table—more about that later.) So the critical region is the set of values that are greater than or equal to 7.8. Because the χ^2 value you calculated for the card problem is only 1.25, it does not fall within the critical region. Thus, you can accept the hypothesis that the deck is honest, or at least, you have no evidence to claim that the deck is dishonest.

Contingency Tables

The above discussion has assumed that you know the probabilities of the possible outcomes of an experiment. Sometimes, though, you have problems to solve in which you don't know the expected probabilities. For example, you might want to study the relationship between level of exposure to a specific air pollutant and the number of abnormalities found in the lungs of exposed laboratory animals. You can collect data on the effects and display them in a contingency table. A **contingency table** is a two-way table that is useful for studying two classification variables. Take a look at Table 13-2. It contains hypothetical data about the effects of air pollutant "z" on 100 lab animals.

Table 13-2. Contingency Table Showing Data On Effects Of Pollutant "z" On Laboratory Mice

Level of exposure to pollutant "z"	Number of lung abnormalities				Totals
	0-2	3-4	5-6	7-8	
High	5	8	17	28	58
Medium	4	4	5	10	23
Low	5	4	5	5	19
Totals	14	16	27	43	100

Remember, when the X^2 value falls outside the critical region, the variation can be attributed to chance and you can accept the hypothesis.

Let's test the hypothesis that there is no relationship between level of exposure and number of lung abnormalities. (Remember, when the X^2 value falls outside the critical region, no relationship exists—that is, the relationship is governed only by chance.) You can't use the X^2 equation you used earlier because you don't know the e_i values—the expected probabilities of each outcome. You can get around that problem, however, by using percentages of totals with observed values to come up with expected probabilities. Suppose you run a number of these experiments in which 58 of the 100 animals are exposed to high levels, 23 of the 100 are exposed to medium levels, and 19 of the 100 are exposed to low levels. You can then use those figures to determine the e_i values under the null hypothesis (see Lesson 11) that there is no relationship between level of exposure and lung abnormalities; thus, the expected number of lung abnormalities in any cell of the table is proportional to the percentage of animals at that exposure level and the percentage of lung abnormalities in that range. It's actually easier to do than say. The percentages in the 0-2, 3-4, 5-6, and 7-8 groups are, respectively, 14, 16, 27, and 43. What you expect in the "high, 0-2" cell is then 58% x 14% x total or $0.58 \times 0.14 \times 100 = 8.12$.

The proportion of animals with high-level exposure is 58%; medium-level, 23%; and low-level, 19%. You can use these percentages to figure all 12 expected frequencies. The expected (e) value in the medium level exposure (0-2) cell is $0.23 \times 0.14 \times 100 = 3.22$. The expected probability in the low, 0-2 cell is $0.19 \times 0.14 \times 100 = 2.66$. Now, figure out the other

values for “e” in the other nine cells in Table 13-3. Then compare your results with Table 13-4.

Table 13-3. Contingency Table With Observed Effects Of Pollutant “z” On Laboratory Mice

Level of exposure to pollutant “z”	Number of lung abnormalities				Totals
	0 - 2 o e	3 - 4 o e	5 - 6 o e	7 - 8 o e	
High	5 8.12	8	17	28	58
Medium	4 3.22	4	5	10	23
Low	5 2.66	4	5	5	19
Totals	14	16	27	43	100

Table 13-4. Contingency Table Showing Data On Effects Of Pollutant “z” On Laboratory Mice; Observed (o) And Expected (e) Frequencies

Level of exposure to pollutant “z”	Number of lung abnormalities				Totals
	0 - 2 o e	3 - 4 o e	5 - 6 o e	7 - 8 o e	
High	5 8.12	8 9.28	17 15.66	28 24.94	58
Medium	4 3.22	4 3.68	5 6.21	10 9.89	23
Low	5 2.66	4 3.04	5 5.13	5 8.17	19
Totals	14	16	27	43	100

Remember the formula for X^2 ?

$$X^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

Use this formula and substitute the 12 pairs of o and e values in Table 13-4. Use your calculator to work these out yourself to make sure you understand how it’s done. Then check your answer with the one worked out on the next page.

$$\begin{aligned}
X^2 &= \frac{(5-8.12)^2}{8.12} + \frac{(8-9.28)^2}{9.28} + \frac{(17-15.66)^2}{15.66} + \frac{(28-24.94)^2}{24.94} \\
&+ \frac{(4-3.22)^2}{3.22} + \frac{(4-3.68)^2}{3.68} + \frac{(5-6.21)^2}{6.21} + \frac{(10-9.89)^2}{9.89} \\
&+ \frac{(5-2.66)^2}{2.66} + \frac{(4-3.04)^2}{3.04} + \frac{(5-5.13)^2}{5.13} + \frac{(5-8.17)^2}{8.17} \\
&= 5.913
\end{aligned}$$

Did you get 5.913? If you didn't, use your calculator and work through the above figures until you can get 5.913. If you did, congratulations! Now, let your finger rest a minute; you just need the degrees of freedom and you'll be almost done.

The number of degrees of freedom in situations like this is calculated using the formula:

Here's the formula for finding the degrees of freedom from values in the table.

$$v = (r - 1)(c - 1)$$

Where: r = the number of rows
 c = the number of columns

For this example, $v = (3 - 1)(4 - 1) = (2)(3) = 6$. Using a X^2 table (Table 13-5), you can see that for $v=6$, the 5% ($P=0.050$) critical value is 12.5916, or rounded off, 12.6. Because your value for X^2 of 5.913 does not fall within the critical region (which is to the right of 12.5916), you can accept the hypothesis that there is no relationship between the level of exposure to pollutant "z" and the number of lung abnormalities found in the laboratory animals.

Table 13-5. Partial Table Of X^2 Distribution Values

$P \backslash v$	0.995	0.975	0.050	0.025	0.010	0.005
1	0.043927	0.039821	3.84146	5.02389	6.63490	7.87944
2	0.010025	0.050636	5.99147	7.37776	9.21034	10.5966
3	0.071721	0.215795	7.81473	9.34840	11.3449	12.8381
4	0.206990	0.484419	9.48773	11.1433	13.2767	14.8602
5	0.411740	0.831211	11.0705	12.8325	15.0863	16.7496
6	0.675727	1.237347	12.5916	14.4494	16.8119	18.5476
7	0.989265	1.68987	14.0671	16.0128	18.4753	20.2777
8	1.344419	2.17973	15.5073	17.5346	20.0902	21.9550
9	1.734926	2.70039	16.9190	19.0228	21.6660	23.5893
10	2.15585	3.24697	18.3070	20.4831	23.2093	25.1882
11	2.60321	3.81575	19.6751	21.9200	24.7250	26.7569
12	3.07382	4.40379	21.0261	23.3367	26.2170	28.2995
13	3.56503	5.00874	22.3621	24.7356	27.6883	29.8194
14	4.07468	5.62872	23.6848	26.1190	29.1413	31.3193

Exercise

You might want to work through a similar problem on your own so you'll feel better about the X^2 test. In this exercise, you'll use the same type of problem, but you'll be given different figures (this one is for pollutant "y"). Using the data in Table 13-6, find the "e" frequencies, then find X^2 . Next, using the X^2 distribution values from Table 13-5, determine whether you can accept the hypothesis that there is no relationship between the level of exposure to pollutant "y" and the number of lung abnormalities found in the laboratory animals.

Table 13-6. Contingency Table Showing Data On Effects Of Pollutant "y" On Laboratory Mice

Level of exposure to pollutant "y"	Number of lung abnormalities				
	0 - 2	3 - 4	5 - 6	7 - 8	Totals
	o e	o e	o e	o e	
High	7	8	10	22	47
Medium	4	6	9	16	35
Low	2	4	5	7	18
Totals	13	18	24	45	100

The solution is on the next page.

Solution

Level of exposure to pollutant "y"	Number of lung abnormalities				Totals
	0 - 2 o e	3 - 4 o e	5 - 6 o e	7 - 8 o e	
High	7 6.11	8 8.46	10 11.28	22 21.15	47
Medium	4 4.55	6 6.3	9 8.4	16 15.75	35
Low	2 2.34	4 3.24	5 4.32	7 8.1	18
Totals	13	18	24	45	100

$$\begin{aligned}
 X^2 &= \frac{(7-6.11)^2}{6.11} + \frac{(8-8.46)^2}{8.46} + \frac{(10-11.28)^2}{11.28} + \frac{(22-21.15)^2}{21.15} \\
 &\quad + \frac{(4-4.55)^2}{4.55} + \frac{(6-6.3)^2}{6.3} + \frac{(9-8.4)^2}{8.4} + \frac{(16-15.75)^2}{15.75} \\
 &\quad + \frac{(2-2.34)^2}{2.34} + \frac{(4-3.24)^2}{3.24} + \frac{(5-4.32)^2}{4.32} + \frac{(7-8.1)^2}{8.1} \\
 &= 0.9458
 \end{aligned}$$

The critical value is 12.5916 (the same as the previous problem), so the X^2 (0.9458) is not in the critical region. Therefore, you can accept the hypothesis there is no relationship between the level of exposure to pollutant "y", and the number of lung abnormalities found in the laboratory animals.